Assignment 3: Data Exploration

Laila Abed

Fall 2024

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Rename this file `<FirstLast>_A03_DataExploration.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Assign a useful **name to each code chunk** and include ample **comments** with your code.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.
7. After Knitting, submit the completed exercise (PDF file) to the dropbox in Canvas.

TIP: If your code extends past the page when knit, tidy your code by manually inserting line breaks.

TIP: If your code fails to knit, check that no `install.packages()` or `View()` commands exist in your code.

Set up your R session

1. Load necessary packages (tidyverse, lubridate, here), check your current working directory and upload two datasets: the ECOTOX neonicotinoid dataset (ECOTOX_Neonicotinoids_Insects_raw.csv) and the Niwot Ridge NEON dataset for litter and woody debris (NEON_NIWO_Litter_massdata_2018-08_raw.csv). Name these datasets "Neonics" and "Litter", respectively. Be sure to include the subcommand to read strings in as factors.

```
#Preparation Code Chunk
#Load packages
library(tidyverse)
library(lubridate)
library(here)
#check work directory
getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```
# Upload and name first database
Neonics <- read.csv(
  file = here('Data','Raw','ECOTOX_Neonicotinoids_Insects_raw.csv'),
  stringsAsFactors = T
)

# Upload and name second database
Litter <- read.csv(
  file = here('Data','Raw','NEON_NIWO_Litter_massdata_2018-08_raw.csv'),
  stringsAsFactors = T
)

#End of preparation code chunk
```

Learn about your system

2. The neonicotinoid dataset was collected from the Environmental Protection Agency's ECOTOX Knowledgebase, a database for ecotoxicology research. Neonicotinoids are a class of insecticides used widely in agriculture. The dataset that has been pulled includes all studies published on insects. Why might we be interested in the ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: I don't have any envrionmental experience, so I serached it online. I found it is very important to understand the ecotoxicology of neonicotinoids for many purposes: 1) Understanding the biodiversity and ecosystem balance including the impact of neonicotinoids on insect species and how it can disrupt the functions of insects and their contribution to decomposition, soil aeration, etc. 2) informing public policies related to environment, 3) health of human and safety of their food, 4) study the impact on bees and other pollinator species population, 5) resistance development in species, 6) assessing risks to various ecosystems.

3. The Niwot Ridge litter and woody debris dataset was collected from the National Ecological Observatory Network, which collectively includes 81 aquatic and terrestrial sites across 20 ecoclimatic domains. 32 of these sites sample forest litter and woody debris, and we will focus on the Niwot Ridge long-term ecological research (LTER) station in Colorado. Why might we be interested in studying litter and woody debris that falls to the ground in forests? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Again based on online search, studying litter and woody debris that falls to the ground in forests could help us understand 1) the nutrient cycling of decomposing and releasing nutrients into the soil, 2) how carbon stored in these materials, released, and affect climate change, 3) the biodiversity where many organisms fungi, insects live in these materials and contribute to the ecosystem, 4) it could impact soil quality and ecosystem functioning. 5) How environment resist or adapt to changes.

4. How is litter and woody debris sampled as part of the NEON network? Read the NEON_Litterfall_UserGuide.pdf document to learn more. List three pieces of salient information about the sampling methods here:

Answer:
1. Litter is collected from elevated traps that are 0.5 m^2 PVC baskets elevated 80 cm above the ground. Fine woody debris is collected from 3 m x 0.5 m rectangular ground traps. 2. Sampling occurs only in tower plots within the 90% flux footprint of primary and secondary airsheds. In forested sites, sampling is in 20-40 40m x 40m plots. In low-stature vegetation sites, it's in 4 40m x 40m plots plus 26 20m x 20m plots. 3. sample frequency varies, where ground traps are sampled once a year. Elevated trap varies: -Deciduous forest sites during senescence (every 2 weeks). -Evergreen sites (every 1-2 months year-round).

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Second code chunk, learn about first data set (Neonics).
#Identify data set dimensions
dim(Neonics)
```

```
## [1] 4623   30
```

```
# 4623 rows, 30 columns
```

6. Using the `summary` function on the "Effect" column, determine the most common effects that are studied. Why might these effects specifically be of interest? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
#Third code chunk, still learn about data set
#Determine most common effects by function summary then sort.
sort(summary(Neonics$Effect))
```

```
##       Hormone(s)        Histology       Physiology          Cell(s)
##                1                5                7                9
##     Biochemistry     Accumulation     Intoxication    Immunological
##               11               12               12               16
##       Morphology           Growth       Enzyme(s)         Genetics
##               22               38               62               82
##        Avoidance      Development     Reproduction Feeding behavior
##              102              136              197              255
##         Behavior       Mortality       Population
##              360             1493             1803
```

Answer:1) Population (impact on the dynamics and abundance of the populations of insects and effects on the biodiversity), 2) Mortality (direct impact of neonicotinoid pesticides on insect population numbers, effects on biodiversity), 3) Behavior (changes and effect on species, resits/adapt, survive, reproduce), 4) Feeding behavior (change/adapt), 5) Repordouction (affect the number of population, biodiversity), 6) Development, 7) Avoidence.

7. Using the `summary` function, determine the six most commonly studied species in the dataset (common name). What do these species have in common, and why might they be of interest over other insects? Feel free to do a brief internet search for more information if needed.[TIP: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```r
#Code chunk know the data set
#Upload package dplyr
library(dplyr)
#Calculate the summary and store it as data frame
Summary_species_common_name <- data.frame(summary(Neonics$Species.Common.Name))
#Sort data frame according to Summary from largest to smallest.
Summary_sorted <- Summary_species_common_name %>%
  arrange(desc(summary.Neonics.Species.Common.Name.))
#Print first 7 rows which are the largest because the first one is
#others (not specific)
head(Summary_sorted,7)
```

```
##                       summary.Neonics.Species.Common.Name.
## (Other)                                                670
## Honey Bee                                              667
## Parasitic Wasp                                         285
## Buff Tailed Bumblebee                                  183
## Carniolan Honey Bee                                    152
## Bumble Bee                                             140
## Italian Honeybee                                       113
```

Answer: Based on online search, the species mentioned have common things: They are pollinators, and contributor to food security especially Honey bees and bumblebees who pollinate fruits, vegetables, and nuts. Also, they have very cooperative social dymnaics and roles which increase human knowledge on communication and cooperation. Moreover, they are indicators of environmental health, biodiversity, and environment resilience through their responses to habitat loss, pesticide exposure, diseases, and climate change. Specially the bees and bumble bees are very interestinhg to study due to declining population.

8. Concentrations are always a numeric value. What is the class of `Conc.1..Author.` column in the dataset, and why is it not numeric? [Tip: Viewing the dataframe may be helpful...]

```r
#Identify the class of a colmun in the data set (Conc.1..Author)
class(Neonics$Conc.1..Author.)
```

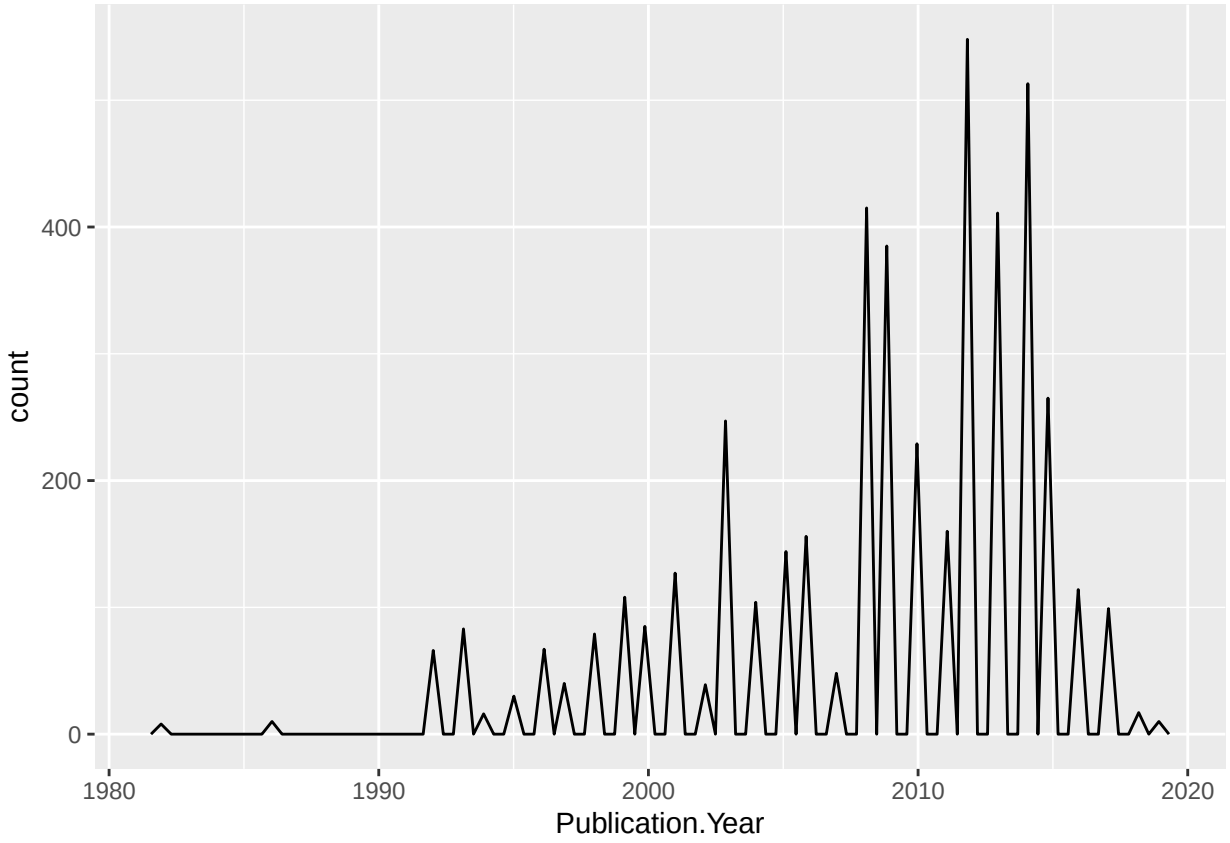
```
## [1] "factor"
```

Answer: The colmun in the data set (Conc.1..Author) contains numbers, symbols and non-numeric characters. So R read it vas string/factor or character.

Explore your data graphically (Neonics)

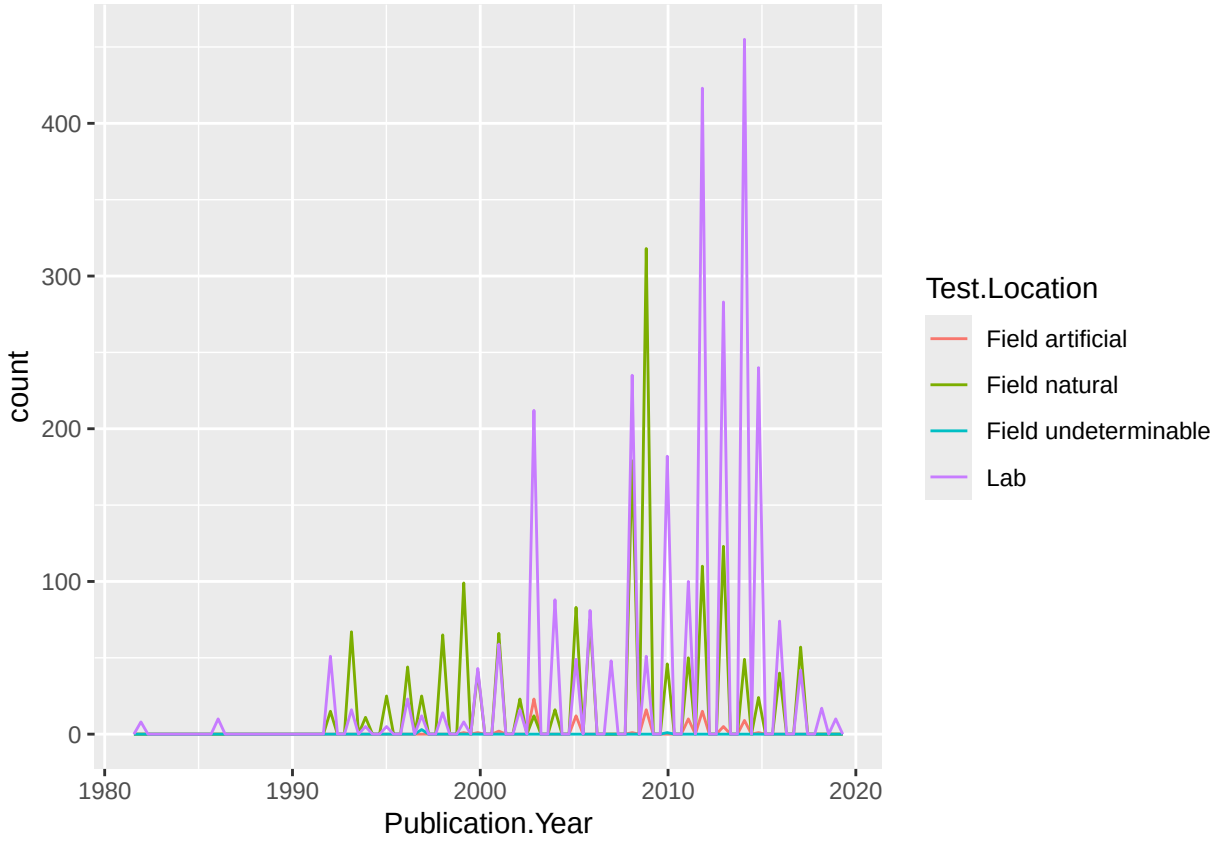
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```r
#Explore data graphically
#Draw a plot of the number of studies conducted by publication year
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 100)
```



10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
#a plot of the number of studies conducted by publication year
#and color different locations
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color = Test.Location), bins=100)
```



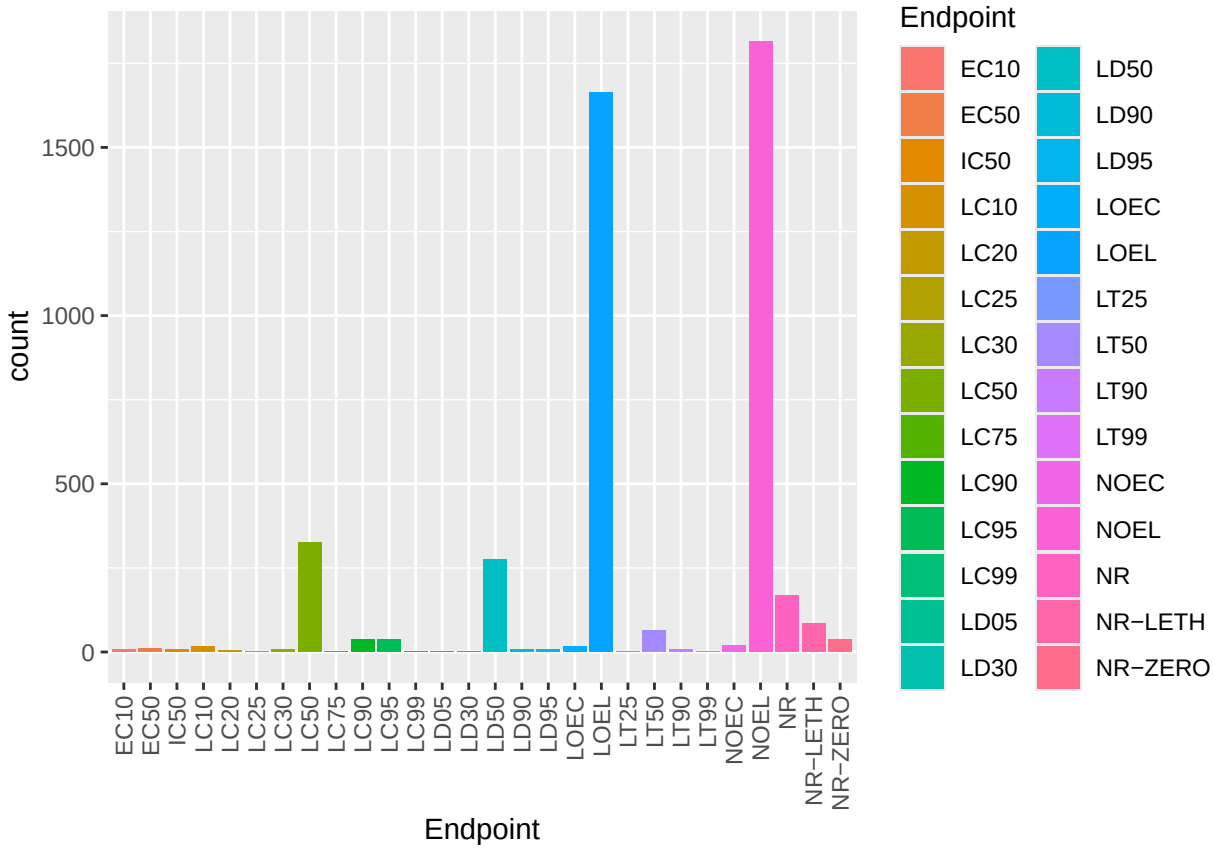
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer:The most common test locations are 1) Lab, we see the number of studies in Lab increase to reach a peak then decreases then increases to reach another peak then decreases. But in general we can consider it incresaes by proceeding with time (year), it reaches its peak around 2015. 2) Field Natural, which reaches its peak around 2010, then the rate declines.

11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

[**TIP**: Add `theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))` to the end of your plot command to rotate and align the X-axis labels...]

```
#Draw a bar graph of Endpoint counts
ggplot(data = Neonics, aes(x = Endpoint, fill=Endpoint)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```
#End of work in the first data set (Neonics)
```

Answer: The two most common endpoints are NOEL and LOEL. NOEL (No Observed Effect Level): is the highest dose or concentration of a substance at which no observable adverse effects or significant changes in the studied organisms are detected during a specific exposure period. NOEL serves as a safety threshold, indicating when exposure to a substance does not result in discernible harm or significant changes in the organisms being studied. LOEL (Lowest Observed Effect Level):is the lowest dose or concentration of a substance at which observable adverse effects or significant changes in the studied organisms are detected during a specific exposure period. LOEL represents the point at which adverse effects become evident, signaling that exposure to the substance is causing harm or significant alterations in the organisms.

Explore your data (Litter)

12. Determine the class of collectDate. Is it a date? If not, change to a date and confirm the new class of the variable. Using the `unique` function, determine which dates litter was sampled in August 2018.

```
#Explore second data set (Litter)
# Classify collectDate
class(Litter$collectDate) #check the class it is not a date, it is factor
```

```
## [1] "factor"
```

```
#change collectdate to date.
Litter$collectDate <- as.Date(Litter$collectDate)
class(Litter$collectDate) #check the class after changing it to date
```

```
## [1] "Date"
```

```
#identify unique values of the date to determine which sampled in August.
unique(Litter$collectDate)
```

```
## [1] "2018-08-02" "2018-08-30"
```

13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
#Determine how many different plots sampled at Niwot Ridge using unique function
#and length.
unique(Litter$plotID) # shows all levels without count of each level.
```

```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```
length(unique(Litter$plotID)) #shows number of levels
```

```
## [1] 12
```

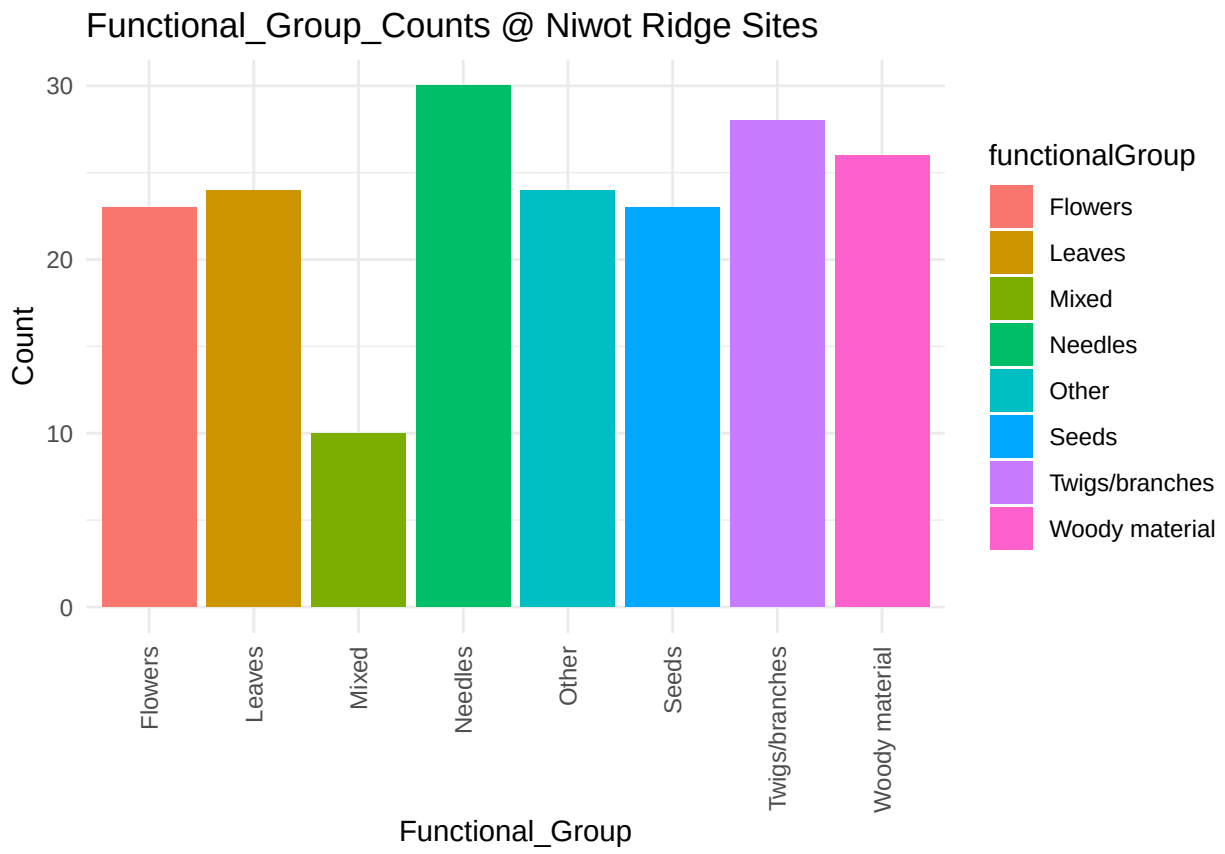
```
summary(Litter$plotID)#shows all levels and count of each one
```

```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: Unique shows and list unique levels without count of each level, it gives indication about total number of levels. Summary shows all levels and count of each one, if the variable is numeric it can do more statistics calculations like mean, median, etc. But if it is a factor like this variable, it shows the count of each level.

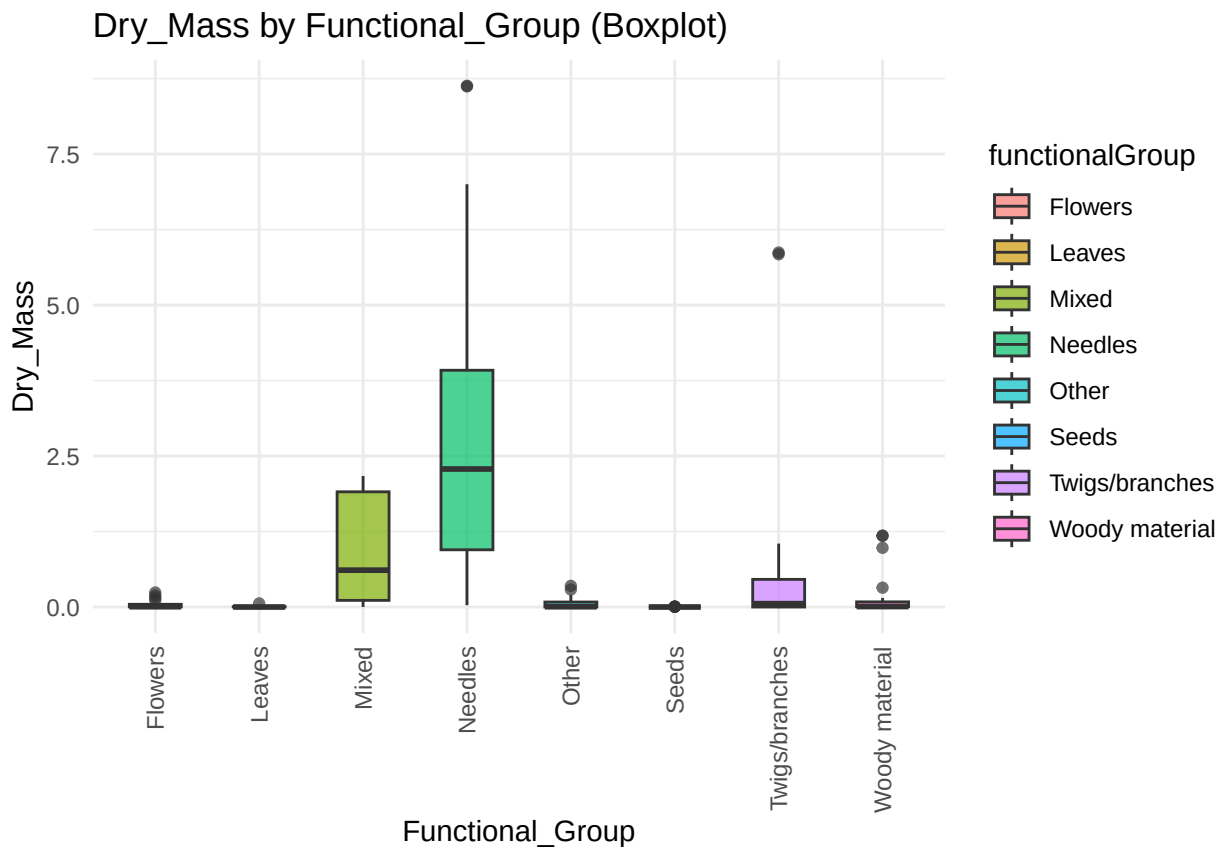
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
#Draw a bar graph of functionalGroup counts.
ggplot(Litter, aes(x = functionalGroup, fill=functionalGroup)) +
geom_bar() +
labs(title = "Functional_Group_Counts @ Niwot Ridge Sites",
x = "Functional_Group",
y = "Count") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

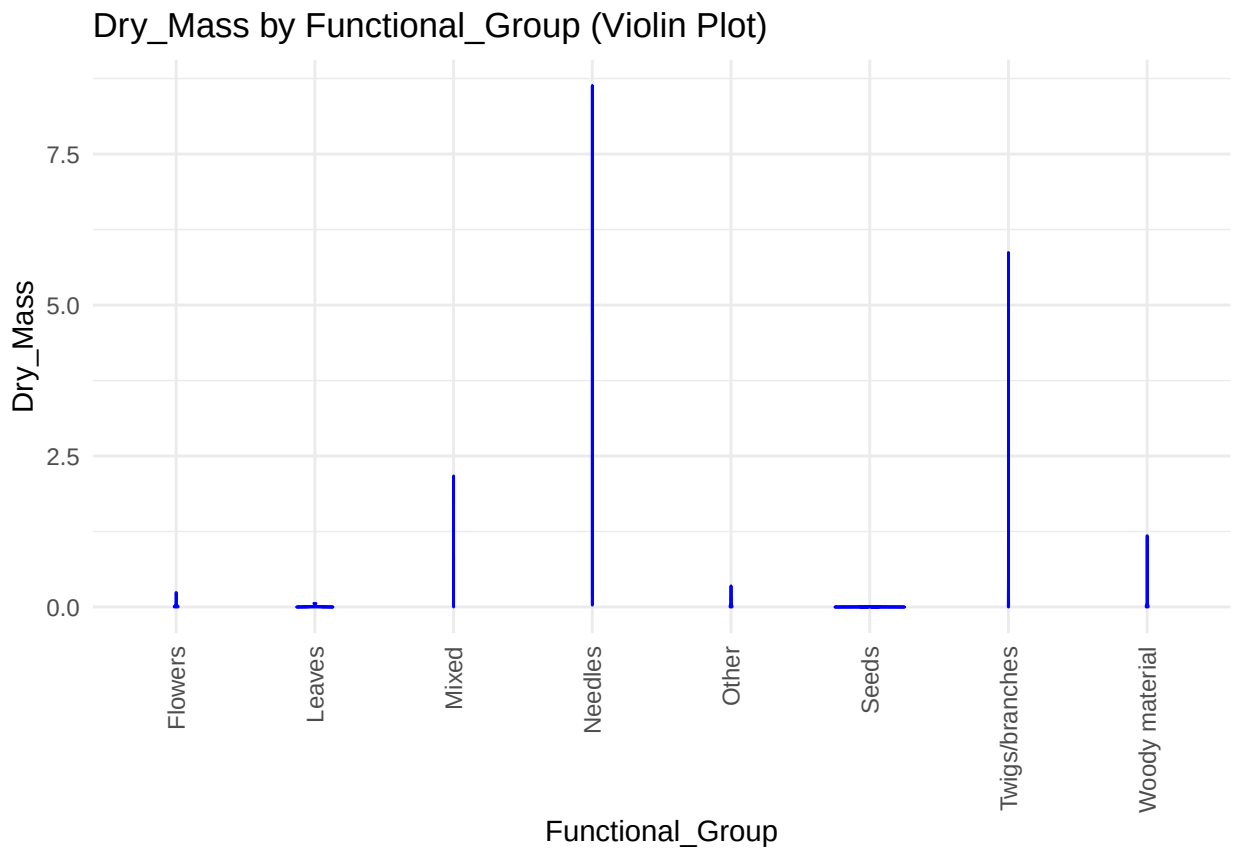


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#Draw boxplot of dryMass by functionalGroup.
ggplot(Litter, aes(x = functionalGroup,fill = functionalGroup, y = dryMass)) +
geom_boxplot(alpha = 0.7, width = 0.5) +
labs(title = "Dry_Mass by Functional_Group (Boxplot)",
x = "Functional_Group",
y = "Dry_Mass") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```




```r
#Draw violin plot of dryMass by functionalGroup.
ggplot(Litter, aes(x = functionalGroup, y = dryMass)) +
geom_violin(color = "blue", alpha = 0.7, width = 0.5) +
labs(title = "Dry_Mass by Functional_Group (Violin Plot)",
x = "Functional_Group",
y = "Dry_Mass") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Why is the boxplot a more effective visualization option than the violin plot in this case?

> Answer: The distribution of dry mass values within each functional group boxplot gives centeral tendency and summary of key statistics like median (horizontal line), the height of each box represents the interquantile range, and outliers (data points outside the range). While, the violin plot shows the median.

What type(s) of litter tend to have the highest biomass at these sites?

> Answer: Needles and Twigs/branches as they have the highest median.